A REVIEW ON PRIVACY PRESERVING DATA MINING TECHNIQUES

K.NagaPrasanthi[1] , Dr.M.V.P.ChandraSekharRao[2]
[1]*Dept. of Computer Science and Engineering, L.B.R.College of Engineering, Mylavaram.*
[2]*Dept.of Computer Science and Engineering, R.V.R&J.C. College of Engineering, Guntur*

***Abstract*— Data mining is a computational process of analyzing and extracting the knowledge from large useful datasets. Most data mining applications operate under the assumption that all the data is available at a single central repository, called a data warehouse. This poses a huge privacy problem because violating only a single repository's security exposes all the data. Privacy preserving data mining techniques are introduced with the aim of extracting the relevant knowledge from the large amount of data while protecting the sensible information at the same time. People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. Several techniques of privacy preserving data mining have been proposed in literature. In this paper, we have studied all these state of art techniques. A tabular comparison of work done by different authors is presented. In our future work we will work on a hybrid of these techniques to preserve the privacy of sensitive data.**

***Keywords*— datamining,privacy preserving, privacy, privacy preserving techniques, sensitive attributes.**

I. INTRODUCTION

Data Mining [1] refers to extracting or "mining" knowledge from large amounts of data. Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. By performing data mining, interesting knowledge, regularities, or high-level information can be extracted from database and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management, and query processing. Data mining is considered one of the most important frontiers in database systems and one of the most promising interdisciplinary developments in the information industry. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. So, there might be a conflict between data mining and privacy. According to the definition, Privacy is the quality or condition of being secluded from the presence or view of others [23]. On relating privacy with data mining, privacy implies keeping information about individual from being available to others [4]. Privacy is a matter of concern because it may have adverse affects on someone's life. Privacy is not violated till one feels his personal information is being used negatively. Once personal information is revealed, one cannot prevent it from being misused. Let us take an example, date of birth, mother's maiden name, or sex etc. may not become a threat for an individual, but if one more attribute like the unique identification number or voter ID are also known then it may cause a serious effect like identity theft. In this paper, we discuss different approaches and techniques in the field of Privacy Preserving Data Mining (PPDM). The paper is organized as follows. In Section 1, we give the basic concept of data mining and privacy. In Section 2, we describe Privacy threats. Section 3 provides classification scheme and evaluation criteria for privacy preserving data mining techniques. Section 4 contains different privacy preserving data mining techniques with their limitations. A tabular comparison of different techniques of PPDM given by different authors is shown in section 5. And finally we conclude in Section 6.

II. PRIVACY THREATS

Releasing the result of data mining could cause privacy threats. Several privacy disclosure threats were possible in micro data publishing like identity disclosure, membership disclosure and an attribute disclosure. Privacy threats results in more disclosure risk. Anonymizing the data and preserving the data through various disclosure protections would result in better utility.

A. Identity Disclosure

Usually an individual was linked to a record in the published table. If his identity was disclosed, then the corresponding sensitive value of an individual would be revealed.

B. Attribute Disclosure

Attribute disclosure was possible when information about individual record would be revealed. Before releasing the data, it is the must to infer attributes of an individual with high confidence. As per the authors view [4], matching multiple bucket was important to protect attribute disclosure.

C. Membership Disclosure

Membership information in the released table would infer an identity of an individual through various attacks. If the selection criteria were not a sensitive attribute value, then it would lead to a membership disclosure [13].

III. CLASSIFICATION SCHEME AND EVALUATION CRITERIA FOR PRIVACY PRESERVING DATA MINING TECHNIQUES

There are various techniques for privacy preserving data mining. Each of the techniques is suitable for particular type of scenario and objectives. We have presented a classification scheme and evaluation criteria for those techniques. However, these schemes and criteria are built based on the classification scheme and evaluation criteria proposed in [2].

A. Classification Scheme

Privacy preserving techniques can be classified based on following characteristics:

- Data Mining Scenario
- Data Mining Tasks
- Data Distribution
- Data Types
- Privacy Definition
- Protection Method

We describe these classification characteristics as follows:

i) Data Mining Scenario: Two major data mining scenario are present basically. In the first scenario, there is no restriction for accessing data and organization releases data for data mining. Data modification is used to achieve the privacy in this scenario. In the second scenario, the organization does not release their data sets but still allow data mining tasks. Cryptographic techniques are basically used for privacy preserving in this scenario.

ii) Data Mining Task: Patterns can be extracted from data sets using various data mining tasks like classification, association rule mining, outlier analysis, clustering and evolution analysis[1]. Generally, the privacy preserving techniques should support all possible data mining tasks and statistical analysis. But they provide support only for some of data mining tasks. Based on this task, the privacy preserving techniques can be categorized.

iii) Data Distribution: Data sets used for data mining can be either distributed or centralized. Phycial location of data is not important but availability/ownership of data is important. Generally the centralized data set is owned by a single party and may be present at computational site or it can be sent to computational site. In case of distributed data set, the data set is shared between two or more parties which need not trust each other private data but interested to perform data mining on joint data. The data set can be homogeneous or heterogeneous. The homogeneous data set obtained by horizontally partitioning data contains same set of attributes but different subset of records. The heterogeneous data set obtained by vertically partitioning data contains different subset of attributes at each site.

iv) Data Types: There are basically two attributes in data set: Numerical and Categorical. Boolean data are the special case of categorical data which takes two possible values 0 and 1. Categorical values lack natural ordering in them. This is the basic difference between categorical and numerical values and it forces the privacy preservation technique to take different approaches for them.

v) Privacy Definition: The definitions of privacy are different in different context. In some scenario individuals' data values are private, whereas in other scenario certain association or classification rules are private. Depending on the privacy definition we work on different privacy preserving techniques.

vi) Protection Methods: Privacy in data mining is protected through different methods such as data modification and secure multiparty computation (SMC). On the basis of protection method we can also categorize the privacy preserving techniques.

B. Evaluation Criteria

It is important to determine the evaluation criteria and related benchmarks. Some evaluation criteria are:

i) Versatility: It refers to the ability of the technique to cater for various data mining tasks, privacy requirements and types of data set. The technique is more useful if it is more versatile. Versatility includes the following:

- Private: Data vs. Patterns
- Data Sets: Distributed or Centralized (Vertical or horizontal)
- Attributes: Numerical or Categorical
- Data Mining Tasks

ii) Disclosure Risks: It refers to the chances of sensitive information being inferred by a malicious data miner. It's inversely proportional to the level of security which is offered by the technique. Development of the disclosure risks is difficult task, since it depends on many factors like supplementary knowledge of an intruder and nature of the technique. Primary objective of privacy preservation technique is minimizing the disclosure risk, so risk evaluation is essential.

iii) Information Loss: Information loss is usually proportional to the amount of noise added and level of security. It is inversely proportional to data quality. The primary requirement of privacy preserving technique is to maintain high data quality in released data sets. If data quality is not maintained then high security will be useless.

iv) Cost: Cost refers to both computation cost and communication cost between the collaborating parties [2]. Computational costs contain both preprocessing cost (e.g., initial perturbation of values) and running cost (e.g., processing overhead). If data set is distributed then communication cost becomes important issue.

IV. PRIVACY PRESERVING DATA MINING TECHNIQUES

There are different PPDM techniques. Some of them are like data perturbation, blocking and cryptography based techniques.

A. Data Perturbation

It is a technique for modifying data using random process[7]. Data distortion or Data noise are the other names for Data Perturbation. This technique can be applied to different data types like character, Boolean, integer and classification type. If X is a set of data records denoted by $X=\{x_1, x_2, ...x_n\}$. For record x_i a noise component is added from probability distribution $f_y(Y)$. These noise components are denoted by $y_1, y_2 ... y_n$. Now the distorted noise records become $x_1+y_1, x_2+y_2, ...x_n+y_n$. The added noise is large enough that original data records can not be guessed from distorted records and only distribution of original records can be recovered.One advantage of Randomization approach is it is relatively simple and does not require knowledge of distribution of other data records. It can be implemented in data collection time itself.

Data distortion or Data noise are the other names for Data Perturbation. Data distortion is done by applying different methods like adding noise, data transpose matrix, adding unknown values etc[3].for every individual problem in classification, clustering or association rule mining, a new distribution based data mining algorithm is to be developed. Another approach in data perturbation is based on singular value decomposition(SVD) and sparsified singular value distribution(SSVD)[3]. Different matrices are introduced to compare or measure the difference between original dataset and distorted dataset. SSVD is efficient in keeping data utility. But there is a drawback in perturbation approach i.e. each data dimension distribution is reconstructed independently. But in many data mining algorithms a lot of relevant information is is hidden in inter-attribute correlations.

B. Blocking Based Technique

Using blocking based technique[4][5], authors state there is a sensitive classification rule for hiding sensitive data from others. There are two steps in this technique.(i)to identify transactions of sensitive rule,(ii)to replace the known values to the unknown values(?). Original database is scanned for identifying transactions supporting sensitive rule. Then for each transaction replace the sensitive data with unknown value. This technique can be applied only if unknown values can be saved for some attributes. Authors[4] state that they have replaced '1' by '0' or '0' by '1'or with any unknown(?) values in a specific transaction. No specific rule is followed while replacing these values. Different types of sensitive rules may exist. When left side of pair of rule is a subset of attribute values pair of transaction and right hand side of rule is same as attribute class of transaction then transaction supports any rule. For any transaction which supports sensitive rule, the algorithm replaces attribute values with unknown values. This will continue until all sensitive attributes are hidden.

C. Cryptographic Technique

Sensitive data can be encrypted using cryptography technique. In [6] , authors introduced cryptographic technique which is very popular because it provides security and safety of sensitive attributes. There are other cryptography algorithms available. But these methods have many disadvantages like they fail to protect the output of computation. The algorithm proposed in[6] does not give fruitful results in case of more parties and also it is very difficult to apply this algorithm to huge databases.

D. Condensation Approach:

Condensation is another privacy preserving approach. *Charu C. Aggarwal and Philip [7]* introduced this method, which builds constrained clusters in the data set and after that produces pseudo-data. Contraction or condensation of data into multiple groups of predefined size is the basic concept of this method. Certain statistics are maintained for each group. This approach is used in dynamic data update such as stream problems. Each group has a size of at least '*k*', which is referred to as the level of that privacy-preserving approach. The higher the level, the higher is the amount of privacy. Statistics from each group are used to generate corresponding pseudo-data. Even though this is a simple privacy preservation approach, it is not efficient one as it leads to loss of information.

E. Hybrid technique:

Privacy preservation is a vast field of research and many algorithms have been proposed in order to secure the data. Two or more techniques can be combined and a new Hybrid technique can be used to preserve the data. *Sativa Lohiya and Lata Ragha* [9] proposed a hybrid technique in which they used randomization and generalization. In this approach, the data is randomized first and then generalization is applied on modified or randomized data. This technique protects data with better accuracy. This technique can reconstruct original data and it provides data with no information loss. A hybrid technique can be constructed by combining many other techniques like Data perturbation, Blocking based method, Cryptographic technique- Condensation approach etc.

V. COMPARISON BETWEEN DIFFERENT TECHNIQUES

There are many different techniques proposed in the field of Privacy Preserving Data Mining but one outperforms over other or vice versa on different criteria. Algorithms are classified on the basis of performance, utility, cost, complexity, tolerance against data mining algorithms etc. We have shown a tabular comparison (table 1) of the work done by different authors in a chronological order (from past to present). We have taken the parameters like technique used for PPDM, its approach, results and accuracy.

TABLE I
TABULAR COMPARISON OF DIFFERENT TECHNIQUES

S. No	Authors	Year of Publication	Technique Used for PPDM	Approach	Result and Accuracy
1.	Y.Lindell, B.Pinkas [11]	2000	Cryptographic Technique	A technique through which sensitive data can be encrypted. There is also a proper toolset for algorithms of cryptography.	This approach is especially difficult to scale when more than a few parties are involved. Also it does not hold good for large databases.
2.	L. Sweeney[22]	2002	K- Anonymity	A record from a dataset cannot be distinguished from at least k-1 records whose data is also in the dataset.	K- Anonymity Approach is able to preserve privacy.
3.	J. Vaidya and C. Clifton[20]	2002	Association Rule	Distribution of data vertically into segments.	Distribution Based Association Rule Data Mining provides privacy.
4.	Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar[7]	2003	Data Perturbation	They tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the "signal" from the data so that the patterns can still be accurately estimated.	Randomization-based Techniques are used to generate random matrices.
5.	CharuC.Aggarwa, Philip S. Yu[12]	2004	Condensation Approach	This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data.	The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data.
6.	A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam [24]	2006	L-Diversity Algorithm	If there are 'l' 'well represented' values for sensitive attribute then that class is said to have L- Diversity.	It is better than K-Anonymity in preserving Data mining.
7.	Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira[21]	2010	Anonymization	Anonymization is a technique for hiding individual's sensitive data from owner's record. K-anonymity is used for generalization and suppression for data hiding.	Background Knowledge and Homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.
8.	P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha[3]	2011	Hybrid Approach	Hybrid Approach is a combination of different techniques which combine to give an integrated result.	It uses Anonymization and suppression to preserve data.
9.	George Mathew, Zoran Obradovic[25]	2011	Decision Tree	An approach which is technical, methodological and should give judgmental knowledge.	A graph-based framework for preserving patient's sensitive information.
10.	Anita Parmar, Udai Pratap Rao, Dhiren R. Patel[10]	2011	Blocking Based Technique	Finding sensitive attribute and then they replace known sensitive values with unknown values ("?"). Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined.	Unknown Values help in preserving privacy but reconstruction of original data set is quite difficult.
11.	Sara Mumtaz, Azhar Rauf and Shah Khusro[16]	2011	Distortion Based Perturbation Technique in OLAP Data Cube	Data perturbation technique which is also called uniformly adjusted distortion is proposed which initially distorts one cell of a cube and then distortion occurs in whole cube.	This distribution of distortion technique not only preserves, but also provides utmost accuracy with range sum queries and high availability.

S. No	Authors	Year of Publication	Technique Used for PPDM	Approach	Result and Accuracy
12.	Hsiang-Cheh Huang, Wai-Chi Fang[17]	2011	Histogram Based Reversible Data Hiding	A concept of reversibility which states that an original data can easily be hidden and the hidden data can also be recovered perfectly. Sensitive data is embedded into medical images which is very good technique for hiding secret data.	Histogram technique is basically used for X-Ray or CT medical images and it has the potential to be integrated into databases for managing the medical images in the hospital.
13.	Jinfei Liu, Jun Luo and Joshua Zhexue Huang[5]	2011	Rating Based Privacy Preservation	A novel algorithm which overcomes the curse of dimensionality and provides privacy	It is better than K-Anonymity and L-Diversity.
14.	Khaled Alotaibi, V. J. Rayward-Smith, Wenjia Wang and Beatriz de la Iglesia[6]	2012	Multi-Dimensional Scaling	A non linear dimensionality reduction technique used to project data on lower dimensional space.	The application of non-metric MDS transformation works efficiently and hence produces better results.
15.	Elahe Ghasemi Komishani and Mahdi Abadi[8]	2012	Trajectory data	Approach for privacy Preservation in trajectory data publishing in which trajectories and sensitive attributes are generalized with respect to different privacy requirements of moving objects.	It is able to provide personalized privacy preservation in trajectory data publishing, but also it is resistant to all three identity linkage, attribute linkage, and similarity attacks.
16.	Thanveer Jahan, Dr. G.Narsimha and Dr. C.V Guru Rao[15]	2012	Data Perturbation Using SSVD	An analyzing system used to transform original dataset into distorted data set using Sparsified Singular Value Decomposition.	Use of Sparsified SVD than SVD is more successful.
17.	D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan[19]	2012	Association Rule	Sanitizes datasets using Sliding Window Algorithm and preserves data.	A novel approach that modifies the database to hide sensitive rules.
18.	M. N. Kumbhar and R. Khara[18]	2012	Association Rule By Horizontal and Vertical Distribution	Different approaches in the field of Association rule are reviewed.	The performance of all models is analyzed in terms of privacy, security and communications.
19.	Savita Lohiya and Lata Ragha[9]	2012	Hybrid Approach	A combination of K- Anonymity and Randomization.	It has a better accuracy and original data can b reconstructed
20.	Martin Beck and Michael Marhöfer[26]	2012	Anonymizing Demonstrator	Making a demonstrator with user friendly interface and performs Anonymization.	Swapping and Recording can be applied to enhance the utility.

VI. CONCLUSION

In today's world, privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don't want to share. Our survey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From our analysis, we have found that there is no single technique that is consistent in all domains. All methods perform in a different way depending on the type of data as well as the type of application or domain. But still from our analysis, we can conclude that Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique for encryption of sensitive data. On the other hand Data Perturbation will help to preserve data and hence sensitivity is maintained .In future, we want to propose a hybrid approach of these techniques.

References

[1] J. Han and M. Kamber , "*Data Mining: Concepts andTechniques*", 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[2] V.S. Verykios, E. Bertino, I. N. Fovino, L. P. Provonza, Y.Saygin and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Record*, 33 (1): 50-57, 2004.

[3] P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in *proceedings of International Conference on Recent Trends in InformationTechnology*, IEEE 2011.

[4] B.Vani, D.Jayanthi, (2013), "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT.

[5] J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in *proceedings of 11th IEEE International Conference on DataMining Workshops*, IEEE 2011.

[6] K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de laIglesia, "Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification" in *proceedings of 2012 ASE/IEEE International Conference on Social Computing and2012 ASE/IEEE International Conference on Privacy,Security,Risk and Trust*, IEEE 2012.

[7] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in *proceedings of the Third IEEE InternationalConference on Data Mining*, IEEE 2003.

[8] E. G. Komishani and M. Abadi, "A Generalization-Based Approach for Personalized Privacy Preservation in Trajectory Data Publishing", *in proceedings of 6'th International Symposium on Telecommunications (IST'2012)*, IEEE 2012.

[9] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in p*roceedings of 2012 FourthInternational Conference on Computational Intelligence and Communication Networks*, IEEE 2012.

[10] A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" ,in *proceedings of International Symposium on Computer Science and Society*, IEEE 2011.

[11] Y. Lindell, B.Pinkas, "Privacy preserving data mining", in *proceedings of Journal of Cryptology*, 5(3), 2000.

[12] C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in *proceedings of International Conference on Extending Database Technology (EDBT)*, pp.183–199, 2004. 746

[13] Tiancheng Li , Jian Zhang , Ian Molloy ,(2012),"Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transaction on KDD.

[14] Evfimievski, A.Srikant, R.Agrawal, and Gehrke , "Privacy preserving mining of association rules", in proceedings of KDD02, pp. 217-228.

[15] T. Jahan, G.Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy" in *proceedings of 978-1-4673-1989-8*/12, IEEE 2012.

[16] S. Mumtaz, A. Rauf and S. Khusro, "*A Distortion Based Technique for Preserving Privacy in OLAP Data Cube*", in proceedings of 978-1-61284-941-6/11/$26.00, IEEE 2011.

[17] H.C. Huang, W.C. Fang, "*Integrity Preservation and Privacy Protection for Medical Images with Histogram-Based Reversible Data Hiding*", in proceedings of 978-1-4577-0422-2/11/ IEEE 2011.

[18] M. N. Kumbhar and R. Kharat, "*Privacy Preserving Mining of Association Rules on horizontally and Vertically PartitionedData: A Review Paper*", in proceedings of 978-1-4673-5116-4/12/$31.00_c, IEEE 2012.

[19] D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan, "A Pattern based framework for privacy preservation through Association rule Mining" in *proceedings of International Conference On Advances In Engineering, Science And Management (ICAESM -2012)*, IEEE 2012.

[20] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", in *The Eighth ACMSIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002*, IEEE 2002.

[21] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira,"Efficient Multi-Dimensional Suppression for K-Anonymity", *inproceedings of IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE2010.

[22] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", in *proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002.

[23] The free dictionary.Homepage on Privacy [Online]. Available: http://www.thefreedictionary.com/privacy.

[24] A. Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond k-Anonymity", *Proc. Int'l Con! Data Eng. (ICDE)*, p. 24, 2006.

[25] G. Mathew, Z. Obradovic," A Privacy-Preserving Framework for Distributed Clinical Decision Support", *in proceedings of 978-1-61284-852-5/11/$26.00 ©2011 IEEE*.

[26] Martin Beck and Michael Marh¨ofer," Privacy-Preserving Data Mining Demonstrator", in *proceedings of 16th International Conference on Intelligence in Next Generation Networks*, IEEE 2012.